AD_____

GRANT NUMBER: DAMD17-94-J-4414

TITLE: cDNA Fingerprinting of Breast Cancer Tumor Cells

PRINCIPAL INVESTIGATOR: Cassandra L. Smith, Ph.D.

CONTRACTING ORGANIZATION: Boston University
Boston, MA 02215

REPORT DATE: October 1997

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Material Command
Fort Detrick, MD 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave Blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1997	Annual Report 30 Sep 96 - 29 Sep 97

4. TITLE AND SUBTITLE
cDNA Fingerprinting of Breast Cancer Tumor Cells

5. FUNDING NUMBERS
Contract DAMD17-94-J-4414

6. AUTHOR(S)
Cassandra L. Smith, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Boston University
Boston, MA 02215

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Material Command
Fort Detrick, MD 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b.

13. ABSTRACT (*Maximum 200 words*)
Several methods have been described that highlight genetic differences between samples. However none of the currently available methods are robust enough to be applied routinely to a large number of samples. The long term goal of this proposal is to develop and to apply such methods to breast cancer disease to aid in diagnosis, prevention and treatment. For instance, it is quite clear that current methods of analysis will be replaced in the future with more efficient and accurate analytically tools such as mass spectrometry (MS). In collaboration with several groups we are developing matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) analysis for DNA analysis. MALDI-TOF MS measurement provides molecular masses with up to 1 Dalton accuracy and 1 part in 1000 resolution in measurements that take less than a second. The improved accuracy of MS analysis will allow the power of array chip technology to be brought on line widely. It is most likely that it will take another two years for the MS technology to mature to the necessary level. Meanwhile our focus has been in developing methods of providing useful samples for the, what will be, sample hungry MS instruments. Rather than wait for the MS technology to mature, these samples are being analyzed now with current but inefficient technology. Our methods have focused on generating pools of samples with common characteristics. The pool of samples is analyzed en masse. For instance, our methods produce pools of samples that contain simple repeating sequences such as (CAG)n and (CA)n repeat sequences, LTR sequences (= a retroviral footprint), or Zn-finger binding motif (= sequences coding for transcription factors). The starting material can be genomic DNA or cDNA. This means that gene expression profiling experiments can focus on sequences important in cancer biology but missed by other methods. Other methods study highly expressed genes whereas our approach can study any type of gene independent of its expression level. Other similar experiments target genomic analysis aimed at detecting tumor cell specific rearrangements, rapid generation of DNA from targeted DNA samples or allow rapid and efficient in situ scoring of genetic markers

14. SUBJECT TERMS	15. NUMBER OF PAGES
breast cancer, cDNA, PCR, chromosome 20, pfg	13
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

__X__ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

__X__ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

__X__ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

__X__ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

__X__ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

A Smith 11/10/97

PI - Signature Date

TABLE OF CONTENTS

Introduction

In positional cloning experiments, conventional genetic methods are used to narrow the search region. Then physical (molecular) methods are applied to further narrow the search region and to identify genes within the search region. Recently, these methods were used to isolate the BCRA1 (Harshman *et al.*, 1995) and BCRA2 (Tavtigian *et al.*, 1996) genes. The Human Genome Project has provided an increasing number of resources for finding disease genes using positional cloning methods. Still, the task of finding genes involved in particular diseases is arduous. Multiple physical methods for identifying genes must be used in each gene search, because no single approach would guarantee the identification of all genes in a particular region.

Thus far the positional genetic approaches have only identified major gene causes. However, the onset or progress of many diseases is governed by multigenic effects and interactions. Even major disease genes are not expressed alone but in a chorus of over 80,000 other genes. Given the spectrum of genomic changes thus far indentified in breast and other cancers, it is quite clear that efficient and reliable methods are needed to analyze the increasing number of genomic sequences important in tumor development, progression and response to therapeutic regimes. Thus, a number of groups, including ours, are focused on developing comparative methods for identifying multi-gene differences between samples that can be applied in a cost effective method to a large number of samples.

Although the published methods for multigene analysis are useful as research tools, none have proven to be robust enough to be routinely applied to samples that have the complexity of the human genome. The approaches include comparative genome hybridization (CGH; Kallioniemi *et al.*, 1994), differential display (Liang and Pardee, 1992; Liang *et al.*, 1994) and subtractive hybridization (Lisitzyn *et al.*, 1993a; Lisitzyn *et al.*, 1993b). In CGH, a mixture of differentially labeled cDNAs from two samples is hybridized to metaphase chromosomes. Genomic regions that are amplified or deleted in one of the test samples will be differentially labeled. Hence, this method can identify genomic regions important in disease states. In differential cDNA display experiments, mRNA levels of appropriate samples are analyzed. Here, total mRNA is amplified randomly and displayed by size, electrophoretically, from different appropriate samples. The differentially expressed cDNAs are then isolated and characterized. In subtractive hybridization, sequences present in one cDNA library but missing in a second cDNA are isolated.

An alternative method of measuring mRNA level is the random sequencing of cDNA libraries made from particular cells. Although several pharmaceutical groups with a large number of resources are taking this approach for some diseases, it is quite clear that DNA sequencing costs at this time preclude the use of this method for routine application. Our original proposal intended to extend the principles of CGH to arrays of cDNAs. Since this proposal was written two methods for differential display of cDNA were described. One method (Schena *et al.*, 1995) is very similar to that described in our original research proposal. The method involves hybridization of differentially labeled cDNA simultaneously to the same array of cDNA probe samples. Schena *et al.* (1995) reported on the application of CGH principles to arrays of yeast cDNAs. We also carried out a number of pilot studies on several arrays of cDNA. The other method (Velculescu *et al.*, 1995) to quantitate gene expression uses direct DNA sequencing of chimeric small clones that are composed of ligated pieces of cDNAs. Each of the ligated pieces is

an index for a particular cDNA. Thus, one sequencing reaction gives information about many cDNAs. The chimeric clones are created in a manner that should preserve quantitative information on the occurrence of each cDNA.

These publications prompted us to rethink our cDNA profiling method. In particular, we are developing a hybrid system using matrix assisted laser desorption/ionization (MALDI) mass spectrometry (MS) as a tool for rapid, cost-effective, comparative studies of cDNAs fragments after specific hybridization capture steps to simplify the mixture of fragments. Recently we and others (Pieles *et al.*, 1993; Roskey *et al.*, 1996) showed that MALDI-MS is an effective tool for the rapid measurement of short (<35 nucleotide) DNA sequences. We have begun to develop the necessary simulation and (data) analytical software tools to adapt MALDI-MS as a method for measuring and characterizing genetic expression. An indexing scheme will be used to identify the cDNA strands corresponding to any given mRNA, or known sequence. Short (n = 8 - 15 nucleotides) sequences are used as identifiers. Such an identifier is capable of identifying 4^n different species. This should provide sufficient indices such that the majority of cDNA species are uniquely represented. Relative percentage abundances of cDNA species must retain those of the corresponding mRNA. The experimental program is investigating methods for maintaining this quantitative information.

Along with the cDNA experiments, we have developed several methods for the MS analysis of genomic DNA. The most advanced, targets important genomic sequences (in a manner similar to that used in the cDNA experiments), and both reduce genomic complexity and focus analysis on regions known or thought to be unstable in tumor cells. Other experiments have focused on developing array technology with genomic DNA directly and with genetic markers. The genomic DNA array is used in place of large clone libraries. This is important because the large genomic clones now in use have rearrangements and deletions. Many of the clones are chimeric; they contain DNAs from different genomic regions. Large insert clone libraries are time consuming and expensive to make and maintain. In contrast, our genomic DNA method can be rapidly applied to any DNA sample. This approach is quite useful in positional cloning searches to access a specific genomic region in a particular DNA sample. Our first application of this approach has been on the q13 region of chromosome 20 known to be amplified in many breast cancer tumor cells. Experiments on *in situ* scoring of simple repeat sequences has focused on improving the methodology and transferring it to solid surfaces.

Body

Novel method development such as the development of MS for DNA analysis is difficult and involves the expertise of collaborators for instumentation, mass spectrometry, chemistry, molecular modeling, engineering, biochemistry, biology etc. The DOA grant monies only pays a portion of the total cost of this program spread over several universities and industry. Our contribution to this collaboration has been developing methods to provide informative samples for analysis. The specific application of the new methods to breast cancer is funded only by this grant.

We have developed a genomic differential display method (Method I) that allows us to compare genomic DNA directly (Broude *et al.*, 1997). The method reduces genome complexity by

capturing genome subsets (i. e. restriction fragments) that contain a targeted interspersed repeat. The captured fragments are labeled with fluorescein and amplified by PCR, then fractionated by size on an automated DNA sequencing instrument. A second method has also been developed that is based solely on PCR (Method II, Broude *et al.*, manuscript in preparation). These methods produce different types of fragments for analysis. Method I produces fragments which contain the target sequence surrounded by unique sequences. Method II produces fragments containing the target sequence at one end of the fragment. The fragments are separated by size so that the display of restriction fragments sizes is obtained.

This past year, our greatest progress has been made with our major approach for generating targeted genomic and cDNA differential display. The methods were first developed on genomic DNA and the target sequence was a simple repeating sequence (CAG)n. The method now can be used with either genomic DNA or cDNA and has now been extended to include other target sequences, a simple repeating sequence (CA)n, an LTR sequence and a sequence coding for a Zn-finger binding motif. The next target sequence will focus analysis on the signaling cascades that are so important in tumor biology. In particular we are currently developing our targeting protocol for classes of G-protein coupled receptors.

The long term objective of this research is to develop simple but accurate methodology that can be used to analyze large informative regions of the genome so that changes at the DNA or RNA levels associated with specific breast cancer characteristics can be uncovered. These changes may occur through point mutations, or larger DNA rearrangments on amplifications. Although a number of similar approaches have been developed and applied to clinical samples, most if not all of the approaches are either quite expensive or too technically demanding to be of wide spread use. Most methods analyze random sequences. This means that when mRNA is studied the sampling will only be on highly expressed genes. In contrast our work has focues on targeted genes that may be expressed at low levels. Many of these types of approaches are plagued by high rates of false positives. Differences between samples are being sought and false positives are false differences between samples. Hence, a great deal of our work has focused on understanding why false positive occurs and how they can be avoided.

The methodology still needs improvement. For instance, we are still exploring the variables that affect the reproducibility of our genomic and cDNA differential display method. These are very tedious experiments that represent an enormous amount of work but absolutely necessary when robut methodology is developed. These experiments involve testing of all of the reaction components against each other in each of the steps to learn the optimum concentrations and incubation times and to learn the error bars allowable on each of variables. We are also continuing our development of methods for automatically analyzing the similarities and differences in our display methods. This will allow us to evaluate different experimental approaches and to determine the level of differences between samples.

We have also begun to apply this methodology to tumor samples. The first samples that we have analyzed were DNAs from a lung sarcoma and normal lung tissues. These samples were choosen because the sample size was large (in contrast to the usual samples availably from breast tumor cells. The results show that our methods can detect different types of fragment polymorphisms using both of our methods (See appendix - Figure 1). This sample set will be extended. We also attempted to analyze and compare some breast cancer tumor cells from paraffin embedded

samples. The DNAs that were provided to us were too degraded to be useful. We are seeking higher quality samples. This may mean that we will need to improve the DNA extraction procedures used for embedded samples. Recently we have also made arrangement to obtain breast cancer biopsy material. This means that although the methodology could still use improvement, we now know enough to apply our method to breast cancer tumor cells.

Our approach to analyzing cDNAs by MALDI-TOF MS is to focus on specific gene classes provided by the methods described above. Hence, we will adopt some indexing technique for sorting the generated targeted fragments to array elements to be analyzed. This combines known and unknown elements in the analysis. A large number of groups are exploring indexing methods. Each method for preparation and selection has its own idiosyncrasies. However, the underlying steps are the same. Generation of an expression profile involves the following:
(1) the creation of cDNA samples using reverse transcriptase,
(2) an index of 10-15 nucleotides within each cDNA is isolated,
(3) PCR amplification of all of the indices is carried out in parallel, and
(4) the relative abundances of the cDNA indices chosen for each cDNA are measured.
The key advantage of indexing is that the PCR amplification is carried out after all of the cDNAs have been reduced to short, same sized DNA fragments. This ought to improve the accuracy of the relative abundance information markedly. The challenge is finding a way to simplify the analysis of the enormous amount of data contained in a full set of indices. Thus, it is quite clear that the number of possible indices and ways of generating them are quite numerous. Thus, we have begun to develop the necessary software tools, simulational and (data) analytical, that are needed for developing and testing the various experimental approaches.

Data reduction will be done through targeting particular sequences and also will be an intrinsic part of the indexing scheme. We will use array hybridization to simplify the mixture of index fragments. Thus, our method in esence combines some features of both indexing as originally suggested by Velculescu *et al.* (1995) with procedures used after more traditional rtPCR as described by Kato (1995, 1996) and Unrau and Deugau (1994). Each index fragment will be generated such that one (single indexing: SI) or both (double indexing:DI) ends have a single-stranded overhang. In each case, one end of the fragment will be hybridized to a spatially separated array of fixed hybridization probes; each probe has a unique single-stranded overhang, and each is analyzed separately by MS. The fixed probe array contains 4^m elements, where m is the number of nucleotides in single-stranded overhang. Our experiments (Broude *et al.*, 1994; Fu *et al.*, 1995) have shown that this greatly reduces the probability of mismatches between the anchored probes and their targets.

Further differentiation of cDNA species is dependent upon whether SI or DI indexing is used (See Appendix - Figure 2). In SI, further differentiation is obtained through mass measurement. In this protocol, only one strand (length N) of the cDNA is analyzed in the MS. Since, m nucleotides are known from the position in the array, this leaves N-m = k nucleotides to be determined by MALDI MS. In a DI approach, a mixture of specifically designed floating probes is hybridized to the second single strand overhang after the cDN fragment has been hybridized into place in the array. For quantitative analysis, competitive hybridization can be used with a mass-labeled set of standards for each array element.

Simulation experiments will guide and optimize the accompanying experimental program which

will be focused on examining the most serious error sources
(1) accuracy of mass measurement by MALDI MS,
(2) hybridization of slightly mismatched probes,
(3) the quantitative represention of mRNAs by the RT-PCR generated cDNAs, and
(4) the coincident occurrence of identical or nearly identical mass labels on different mRNA species.

These modeling experiments will also take advantage of the the National Cancer Institute's, Cancer Genome Anatomy Progect to include the ever increasing number of genes that have been identified to play some role in breast and other cancers. Eventually, these genes will make up another of our MS test systems since differential display has already been used to assess the level of these genes in about 20 different breast cancer cell lines and primary tumor cells.

Genes and other important sequences which may be important in breast cancer can be fished from particular genomic regions. This is important because position cloning experiments identify such regions. Then a great deal of effort is made at analyzing the regions. Here, we have used genomic DNA directly from a region of human chromosome 20 amplified in breast cancer tumor cells. This region was identified by CGH experiments of others (Tanner, *et al.*, 1994) who have then used time consuming conventional positional cloning approaches to identify putative genes important in breast cancer.

Our approach uses pulsed field gel- (PFG: Schwartz *et al.*, 1983; Schwartz and Cantor, 1994) fractionated genomic restriction fragments as a direct source of DNA (Mass *et al.*, manuscript in preparation). Genomic DNA that has been cut with a restriction enzyme is fractionated by PFG under appropriate conditions. The gel lane containing DNA is cut into 2 mm slices. Each slice is melted in a solution containing 20 mM of ethanolamine by heating to 95° C 15 min. These samples can be stored indefinitely. The DNA in agarose can be used as a template in a number of reactions including PCR. For instance, PCR reaction can be used to test for the presence of particular STS's in slices.

We have used the DNA contained in slices to analyze a region of chromosome 20 amplified in breast cancer tumor cells. The experiments used genomic DNA from a monosomic hybrid cell line containing human chromosome 20. STS analysis of 22 sequences identified slices containing DNA from the amplified region. Then, long inter-*Alu* PCR was used to amplify and ^{32}P-label human DNA from the amplified region. The labeled DNA was used as a hybridization probe to screen a heterogeneous nuclear (hn)cDNA library. About ninety clones were identified that hybridized to this region. Other available genomic resources (e. g. cloned sequences) were also used as hybridization probes. Eight clones with high intensity hybridization signals were sequenced. Then, STS PCR primers were designed, and gel slices and available large insert clones in the amplified region were tested for the occurrence of the selected sequences. The results of these experiments indicate that the majority of these test clones come from the selected chromosomal region. This confirms other experiments done in collaboration with Joe Gray using FISH (flourescent *in situ* hybridization) that demonstrated that our gel slices provided region - specific DNA. This past year we have explored the best way of amplifying the genomic DNA in the slices so that the template DNA supply from a single experiment can be used in many applications. Eventually, our goal would be to use such slices as an array target for experiments similar to, but easier than, CGH.

Summary

Great technical progress has been made with our developing methods. Several articles are being written up now which focus mostly on the methodology. However, we have now begun to apply the methods to a small number of breast cancer tumor cells to identify the problems that are posed by the pecularities of those samples

The major progress on genomic profiling entails the realization that the originally proposed method is not as powerful as newly developing MS methods. Thus, we decided to take a very forward looking approach to cDNA profiling, rather than use the current inefficient methods. Fortunately, the basic methods of DNA handling are almost the same as those proposed in the original grant. Specific adaption to MS is now being done. Meanwhile several other methods for speeding gene searches have also been developed.

References

Brenner, S. and Livak, K. J. 1989. Proc. Nat. Acad. Sci., 86: 8902-8906.

Broude, N., Sano, T., Smith, C. L., and Cantor, C. R. 1994. Proc. Nat. Acad. Sci. USA 91: 3072-3076.

Broude, N., Chandra, A., and C. L. Smith. 1997. Proc. Nat. Acad Sci. USA 94: 4548-4553.

Broude, N., Lykyanov, S. A., Graber, J. H., Sverdlov, E. D., and Smith, C.L. 1997. Submitted.

Ermolaeva, O. D. and Sverdlov, E. D. 1995. Genetic Analysis (Biomolecular Engineering) 13: 49-54.

Fu, D. J., Broude, N. E., Koester, H, Smith, C. L., and Cantor, C. R. 1995. Genetic Analysis (Biomolecular Engineering) 12: 137-142.

Harshman, K., Bell, R., Rosenthal., J., *et al.*, 1995. Hum. Mol. Genet. 4: 1259-1266.

Kallioniemi, A., Kallioniemi, O.-P., Piper, J., Tanner, M., Stokke, T., Chen, L., Smith, H. S., Pinkel, D., Gray, J. W., and Waldman, F. M. 1994. Proc. Natl. Acad. Sci. USA 91: 2156-2160.

Kato, K. 1995. Nucleic Acids Res. 23: 3685-3690.

Kato, K. 1996. Nucleic Acids Res. 24: 394-395

Liang, P. and A. B. Pardee 1992. Science 257: 267

Liang, P., Zhu, W., Zhang, X., Guo, Z., O'Connell, Averbukh, L., Wang, F. and Pardee, A. 1994. Nucleic Acids REs. 22, 5763-5764.

Lisitzyn, N., Lisityn, N., and Wigler M. 1993a. Science 259, 946-951.

Lisitzyn, N., Rosenberg, M., Launer, G., Wagner, L., Potapov, V., Kolesnik, T., and Sverdlov, E. 1993b. Mol. Genet. Microbiol. and Virusol. 3, 26-29.

Sager, R, Sheng, S., Anisowixz, A., Sotiropouloou, G., Zou, A., Stenman, G., Swisshelm, Z., Chen, M., Hendrix, J.C., Pemberton, P., Rafidi, K., and Ryan, K. 1994. Cold Spring Harbor Symposia on Quant. Bio. LIX: 537-545.

Schena, M., Shalon, D., Davis, R. W., and Brown, P.O. 1995. Science 270, 467-70

Smith, D. R. 1992. PCR Meth. Appl., 2: 21-27

Tanner, M., M. Tikkonen, A Kallioniemi, C. Collins, T. Stokke, R. Karhu, D. Kowbel, F. Shadravan, M. HIntz, W.L. Kuo, f. M. Waldman, J. Isola and J. W. Gray. 1994. Cancer Research 54: 4257-4260.

Tavtigian, S. V., Simard, J., Rommens, J., *et al.* 1996. Nat. Genet. 12: 333-337.

Unrau, P. and Deugau, K. V. 1994. Gene 145: 163-169.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. 1995. Science 270: 484-487.

Appendices

Figure 1. Comparison of (CAG)n containing genomic restriction fragments from sarcoma and normal cells. The intensity (y-axis) vs size (x-axis) provides information on the size distribution of HaeIII fragments containing the targeted (CAG)n sequence.

Figure 2. SI and DI Approaches to MALDI MS cDNA profiling.

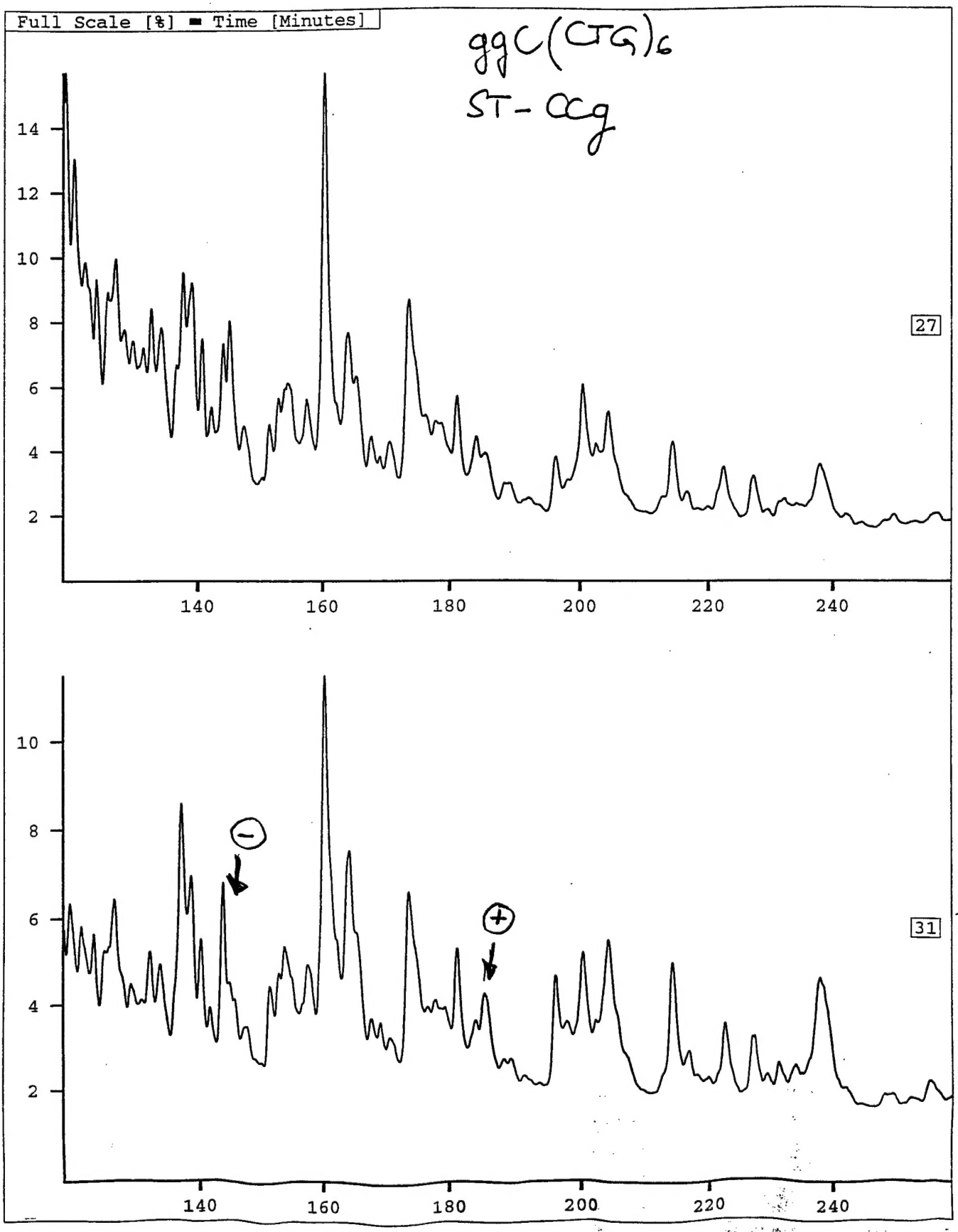
Full Scale [%] ■ Time [Minutes]
ggC(CTG)6
ST-Ccg
14
12
10
8
6
4
2
140
160
180
200
220
240
27
N
10
8
6
4
2
140
160
180
200
220
240
31
T

Figure 1

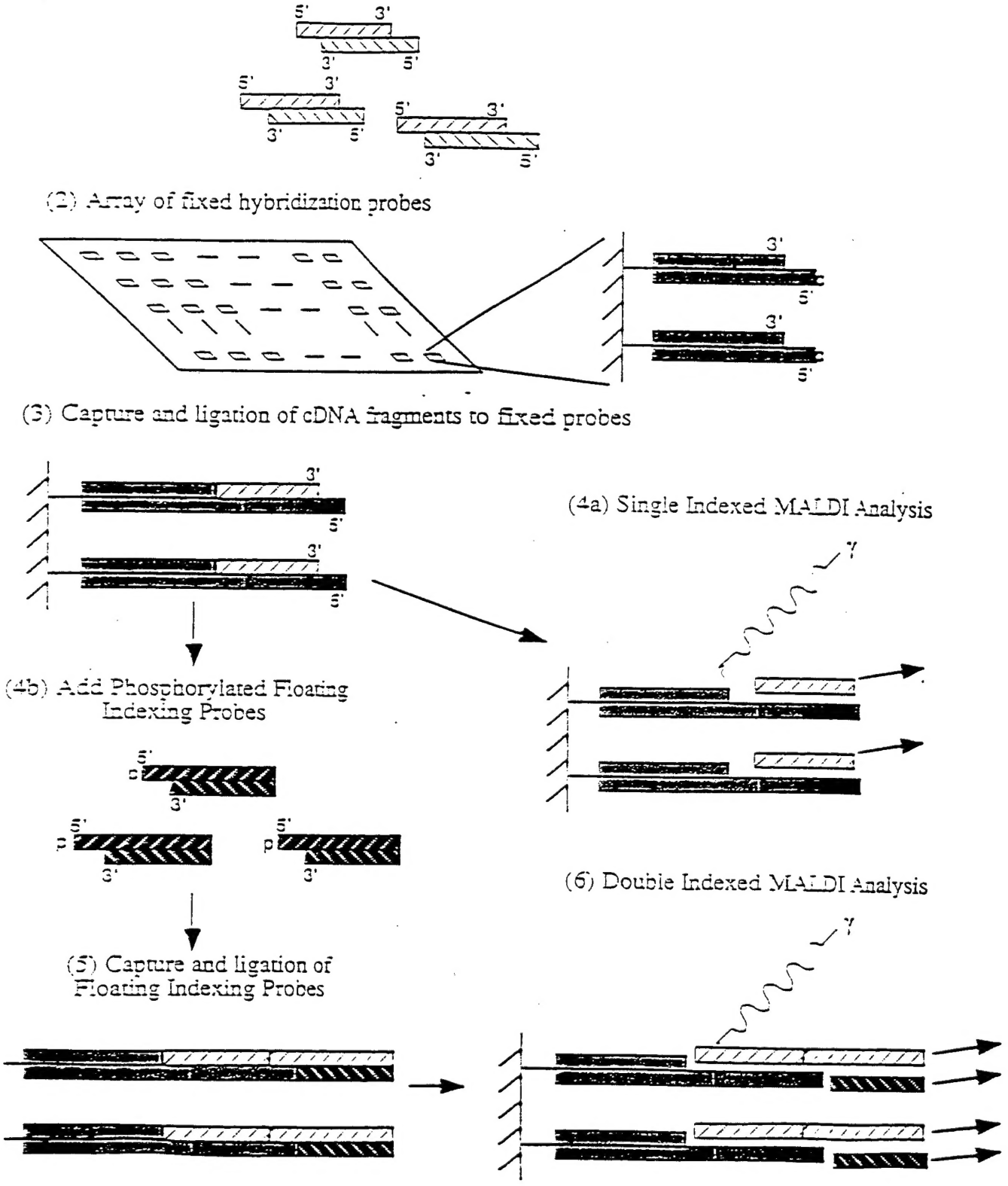

Figure 2 Graphical representation of MALDI measurement of genetic expression. in both the single indexing and double indexing schemes. Pattern changes in fragments are used to show the joining together of fragments through ligation.